SPEAKER AGE ESTIMATION ON CONVERSATIONAL TELEPHONE SPEECH USING SENONE POSTERIOR BASED I-VECTORS

Seyed Omid Sadjadi, Sriram Ganapathy, Jason W. Pelecanos

IBM Research
Yorktown Heights, NY, USA
{sadjadi,ganapath,jwpeleca}@us.ibm.com

ABSTRACT

Automatic age estimation from speech has a variety of applications including natural human-computer interaction, targeted advertising, customer-agent pairing in call centers, and forensics, to mention a few. Recently, the use of i-vectors has shown promise for automatic age estimation. In this paper, we adopt a phonetically-aware i-vector extractor for the age estimation problem. Such senone i-vector based schemes have demonstrated success in the speaker recognition field. Fixed-length and low-dimensional i-vectors are first conditioned through a linear discriminant analysis (LDA) transform, and then used to train a support vector regression (SVR) model. Additionally, in contrast to previous work, we employ the use of the logarithm of the age as the target in training the SVR to further penalize estimation errors for younger speakers compared with older speakers. The proposed system is evaluated using telephony speech material extracted from the NIST SRE 2008 and 2010 evaluation corpora. Experimental results indicate solid age estimation performance with a mean absolute error (MAE) of 4.7 years for both male and female speakers on the NIST SRE 2010 telephony test set.

Index Terms— Age estimation, deep neural networks, i-vector, linear discriminant analysis, support vector regression

1. INTRODUCTION

Speech is a unique physiological signal that contains information at multiple levels regarding the linguistic content (such as words, message, accent, language) as well as the paralinguistic content (such as gender, age, identity, emotional state). It also carries useful information regarding the acoustic conditions of the environments through which it is produced and transmitted (e.g., ambient noise, transmission channel). With the proliferation of mobile devices which enable seamless remote speech acquisition (e.g., over the Internet), there is a growing demand for the development of audio analytics tools to not only extract the message, but also gain insights into the paralinguistic content. Automatic extraction of such speaker/user dependent content from speech has a wide range of applications including natural interaction with dialogue systems, caller-agent pairing in call-centers, user characterization, age targeted advertising, and forensics (see [1] for a comprehensive list of applications). In this study, we focus on automatic speaker age estimation from telephone speech.

Several techniques have been previously proposed in the literature to either classify speech samples into broad age categories (e.g., child, young, adult, and senior) [1, 2, 3, 4], or compute an exact number as the age estimate [5, 6, 7, 8]. These techniques can be categorized into 1) feature based methods [3, 4, 9, 10] where the focus is on identifying a robust feature subset (or an early/late combination of multiple features) that can capture more accurately the age information from speech using standard classification/regression algorithms, and 2) back-end based methods [3, 5, 8, 11, 12] where the goal is to either develop or identify a classification/regression algorithm that can effectively estimate the age information from standard speech representations such as the mel-frequency cepstral coefficients (MFCC) [13]. Recently, the use of i-vectors [14] along with algorithms such as support vector regression (SVR) and artificial neural networks (ANN) have shown great promise for the task of automatic age estimation [6, 7, 8]. It was, however, reported in [8] that when the i-vector representation of speech is used as input feature, the choice of back-end did not seem to significantly impact the age estimation results.

Accordingly, in this study we use a standard SVR back-end and focus on improving the front-end i-vector representation. More specifically, motivated by large improvements seen with the senone posterior based i-vectors for speaker recognition [15], we use a deep neural network (DNN) acoustic model (as opposed to a Gaussian mixture model - GMM) to compute the frame-level soft alignments required in the i-vector estimation process. Additionally, in contrast to previous work, i) we show that the application of linear discriminant analysis (LDA) on i-vectors can improve the computational efficiency as well as help select directions more relevant for the age estimation task, and ii) we employ the use of the logarithm of the age as the target in training the SVR model to emphasize relative (as opposed to absolute) regression errors, thereby further penalizing estimation errors for younger speakers compared with older speakers. Experimental results indicate that the proposed system, which is trained on English conversational telephone speech (CTS) material extracted from the 2008 NIST speaker recognition evaluation (SRE) data, achieves a mean absolute error (MAE) of 4.7 years for both male and female speakers on the NIST SRE 2010 telephony test set.

2. AGE ESTIMATION SYSTEM

In the following subsections, we briefly describe the major components of the age estimation system proposed in this study. We also elaborate on what distinguishes the current system from the previous work in [6, 7, 8]. A schematic block diagram of the system is depicted in Fig. 1.

2.1. I-vector feature extraction

Motivated by the promising outcomes of the recent work on age estimation [6, 7, 8], in this study we also use i-vectors to represent speech samples. The i-vector representation is based on the total

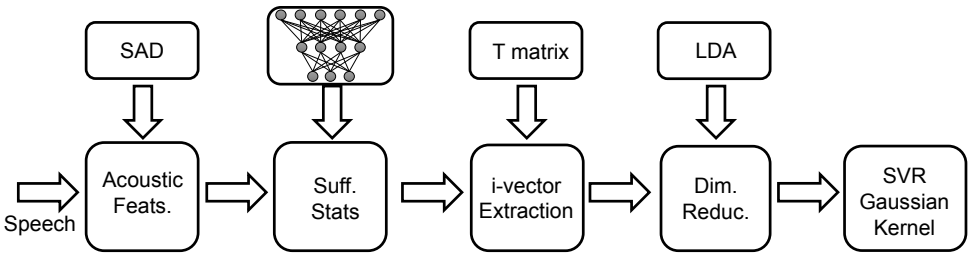

Fig. 1. Block diagram of the age estimation system with DNN senone posterior i-vectors and dimensionality reduction.

variability modeling concept which assumes that speaker- (i.e., identity, age, language) and channel-dependent variabilities reside in the same low-dimensional subspace [14]. The key idea here is that variability within and across sessions can be described via a small set of parameters (a.k.a factors) in a low-dimensional subspace spanned by the columns of a low-rank rectangular matrix, $\mathbf{T}$, entitled the *total variability matrix*. Mathematically, the adapted mean supervector, $\mathbf{M}(s)$, for a given set of observations, s, can be modeled as,

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{T}\,\mathbf{x}(s) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{m}$ is the prior mean supervector, $\mathbf{x}(s) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a multivariate random variable termed an identity vector "i-vector", and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is a residual noise term to account for the variability not captured via $\mathbf{T}$. In other words, for the given observation set s, the i-vector represents the coordinates in the total variability subspace.

In order to learn the bases for the total variability subspace, one needs to compute the Baum-Welch statistics which are defined as,

$$\begin{aligned} N_k(s) &= \sum_t \gamma_{tk}(s), \\ \mathbf{F}_k(s) &= \sum_t \gamma_{tk}(s)\,\mathbf{O}_t(s), \end{aligned}$$

where $N_k(s)$ and $\mathbf{F}_k(s)$ denote the zeroth- and first-order statistics for speech session s, respectively, with $\gamma_{tk}(s)$ being the posterior probability of the mixture component k given the observation vector $\mathbf{O}_t(s)$ at time frame t. Traditionally, $\gamma_{tk}(s)$ is computed with a GMM acoustic model trained in an unsupervised fashion (i.e., with no phonetic labels). However, in [16], a supervised GMM-HMM acoustic model (derived from a speech recognition system) was utilized to estimate the GMM-UBM hyperparameters for speaker recognition, assuming that class-conditional distributions for the various phonetic classes are Gaussian. More recently, inspired by the success of DNN acoustic models in automatic speech recognition (ASR) field, [15] proposed the use of DNN senone (context-dependent triphones) posteriors for computing the soft alignments, $\gamma_{tk}(s)$, which resulted in remarkable reductions in speaker recognition error rates. Motivated by these results, in this study, we explore the senone posterior based i-vectors for the age estimation task, and compare their effectiveness against GMM i-vectors on this task.

2.2. Linear discriminant analysis (LDA)

As noted before, i-vectors model speaker- and channel-dependent information within the same total variability subspace. Therefore, in order to select the most relevant feature subset for the age estimation task, LDA can be applied to i-vectors to annihilate the directions not informative for age estimation. In addition, reducing the dimensionality of i-vectors via LDA can improve the computational efficiency of the subsequent components in the system.

LDA computes an optimum linear projection $\mathbf{A}: \mathbb{R}^d \mapsto \mathbb{R}^n$, by maximizing the ratio of the inter-class scatter to intra-class variance, where $\mathbf{A}$ is a rectangular matrix with n linearly independent columns. Here, the within- and between-class scatter matrices are used to formulate a class separability criterion which converts the matrices into a single statistic. This statistic takes on larger values when the between-class scatter is larger and the within-class variance is smaller. Several such class separability criteria are described in [17], of which the following is the most widely used,

$$\hat{\mathbf{A}} = \underset{\mathbf{A}^T\mathbf{S}_w\mathbf{A}=\mathbf{I}}{\arg\max} \left[\mathrm{tr}\left(\mathbf{A}^T\mathbf{S}_b\mathbf{A}\right)\right], \tag{2}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ denote the between- and within- class scatter matrices, respectively. The optimization problem in (2) has an analytical solution that is a matrix whose columns are the n eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

In this study, an LDA transform is learned utilizing the same training data used for age estimation with discrete chronological age labels (in years). This is in contrast to the previous work in [6, 7, 8] that used background data with speaker labels to train the LDA transform. It will be shown in Section 4 that post-processing i-vectors in this manner not only helps reduce the computational complexity for the SVR model, but also improves the accuracy of age estimation.

2.3. Support vector regression (SVR)

Recently, the use of i-vectors along with SVR has achieved success for the task of automatic age estimation [6, 7, 8]. SVR [18] is a function estimation algorithm that works by constructing an optimal regression hyperplane, which has at most ϵ deviation from the true targets for most training examples (assuming a soft-margin scenario), and at the same time is of minimum norm. More precisely, let $\mathcal{D} = \left\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}\right\}_{i=1}^N$ denote a training dataset in which each input example $\mathbf{x}_i$ is associated with a true target y_i. Here, the goal is to find a regression function $f(\mathbf{x}_i)$ that best describes the input-output relationship, $(\mathbf{x}_i, y_i)$, over the entire training set. Solving the dual optimization problem for the regression hyperplane results in an SVR model $f(\mathbf{x})$ which has the following form,

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b, \tag{3}$$

where $k(\cdot,\cdot)$ denotes a kernel function which must be positive semidefinite [19], $\alpha_i \in \mathbb{R}$ is a Lagrange multiplier, and $b \in \mathbb{R}$ is a bias. In this study, we use a Gaussian kernel function for SVR.

It is worth noting that the optimization constraints applied in obtaining the SVR model in (3) only take into account the absolute difference between the estimated and actual targets. While this may be an appropriate criterion for many applications, for the age estimation task emphasizing relative (as opposed to absolute) regression errors seems more meaningful. For instance, an estimation error of 7 years

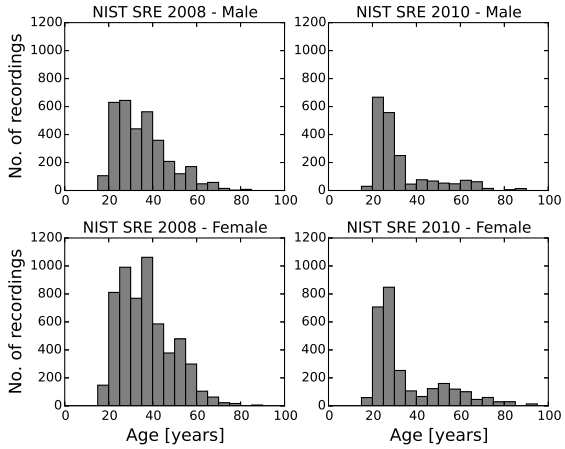

Fig. 2. Age distributions of speakers in the training and test sets extracted from NIST SRE 2008 and 2010 corpora, respectively.

for a 20 year-old speaker should incur a much larger loss compared to that for an 80-year old speaker. Accordingly, in this study we employ the use of the logarithm of the age as the target in training the SVR to further penalize estimation errors for younger speakers compared with older speakers. More specifically, in the training phase the original target ages are transformed as,

$$g(y) = \ln(y - \beta), \tag{4}$$

where $\ln(\cdot)$ denotes the natural logarithm, and $\beta \in \mathbb{R}$ is a task-dependent offset parameter that can be adjusted to achieve the desired level of emphasis on the relative regression errors. In this study, we set the offset parameter as $\beta = y^{\min} - \delta$, where $y^{\min}$ is the minimum age in the training set, and $\delta \in \mathbb{R}^+$ is a small constant to avoid the logarithm of zero. In the test phase, we apply the inverse function $g^{-1}(.)$ to the output of the SVR model to recover the age estimate. Notice that the offset parameter β also has the benefit of setting a lower bound on the estimated ages.

3. EXPERIMENTS

This section provides a description of our experimental setup including speech data, the age estimation system configuration, as well as the performance metrics used in our evaluations.

3.1. Data

We conduct the core of our age estimation experiments using CTS material extracted from data corpora released through the linguistic data consortium (LDC) for the NIST SRE 2004-2010 [20, 21]. These copora contain speech data spoken in U.S. English from a large number of speakers with multiple sessions per speaker. The NIST SRE 2008 and 2010 data also include rich meta data regarding speakers' place of birth, age, height, weight, etc, making them applicable for age estimation experiments. Speech recordings from the *short2-short3* core condition in the NIST SRE 2008 data are utilized for training a gender-independent SVR model, while speech data from the NIST SRE 2010 telephony core condition are used as test material. It is worth noting that there is no overlap between the two corpora (neither speakers nor recordings). Fig. 2 shows the age histograms of male (first row) and female (second row) speakers in the training and test sets extracted from the NIST SRE 2008 and 2010 corpora, respectively.

3.2. System configuration

For speech parameterization, we extract 20-dimensional MFCCs (including c_0) from 25 ms frames every 10 ms using a 24-channel mel filterbank spanning the frequency range 125-3800 Hz. The first and second temporal cepstral derivatives are also computed over a 5-frame window and appended to the static features to capture the dynamic pattern of speech over time. This results in 60-dimensional feature vectors. We also explore shifted delta cepstra (SDC) features [22, 23] with the commonly used 7-1-3-7 (N-d-P-k) configuration. The SDC features are appended to the static cepstral coefficients ($c_0 \cdots c_6$) resulting in a 56-dimensional feature vector. For non-speech frame dropping, we employ an unsupervised speech activity detector (SAD) based on voicing energy features [24]. After dropping the non-speech frames, global (recording level) cepstral mean and variance normalization (CMVN) is applied to suppress the short term linear channel effects.

In this study, a 500-dimensional total variability subspace is learned and used to extract i-vectors from the age estimation training and test sets. To learn the i-vector extractor, we select a total of 13,776 English telephone recordings (from 600 male and 823 female speakers) from the NIST SRE 2004, 2005, and 2006 corpora. The zeroth and first order Baum-Welch statistics are computed for each recording using soft alignments obtained from either a gender-independent 1024-component GMM-UBM with diagonal covariance matrices, or a DNN acoustic model with 2,451 softmax output units that correspond to senones. The DNN, which has 7 hidden layers with 2048 units per layer, is discriminatively trained using the standard error back-propagation and cross-entropy objective function to estimate posterior probabilities of the senones which are obtained by merging 10,000 HMM triphone states using a decision tree with maximum-likelihood (ML) criterion [25]. The DNN training is accomplished on 600 hours of CTS data from the Fisher corpus [26] using a 9-frame context of 40-dimensional speaker-adapted feature space maximum likelihood linear regression (fMLLR) [27, 28] features generated with alignments obtained from a GMM-HMM acoustic model (see [29] for more details).

Prior to training the SVR, for the sake of feature selection and dimensionality reduction, the i-vectors are processed through an LDA transform trained using the NIST SRE 2008 training set with 62 discrete age categories between 16 to 84 years. This is followed by scaling (normalizing) the features to [-1, 1] range. A gender-independent SVR model is then trained on the dimensionality reduced and normalized i-vectors. To alleviate the impact of data imbalance in the NIST SRE 2008 training set (see the first column of histograms in Fig. 2), in particular for speakers older than 50 years, we employ a per-sample weighting scheme to force the regression model to place more emphasis on these points. In particular, the weights for the age bracket $y_i \geq 50$ are set to 5.0 (versus 1.0 for the rest of data) to account for the sparseness of the data in that age bracket.

3.3. Evaluation metrics

To evaluate the age estimation performance, we use two commonly adopted objective measures: i) the mean absolute error (MAE), and ii) Pearson's correlation coefficient. The MAE is defined as,

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^{M} |\hat{y}_i - y_i|,$$

where y_i and $\hat{y}_i$ denote the true and estimated ages for the i^{th} test sample, and M is the number of examples in the test set. A smaller value on the MAE measure indicates better performance. The sample Pearson correlation coefficient between the true and estimated ages is computed as,

Table 1. Age estimation performance with and without LDA.

	Male		Female		Both	
System	MAE	ρ	MAE	ρ	MAE	ρ
MFCC - w/o LDA	6.6	0.79	6.1	0.85	6.3	0.83
MFCC - w/ LDA	**6.5**	0.79	**5.7**	**0.87**	**6.0**	**0.84**

Table 2. Age estimation performance with and without logarithmic optimization constraints.

	Male		Female		Both	
System	MAE	ρ	MAE	ρ	MAE	ρ
MFCC - w/o log	6.8	0.76	6.3	0.86	6.5	0.83
MFCC - w/ log	**6.5**	**0.79**	**5.7**	**0.87**	**6.0**	**0.84**

$$\rho = \frac{1}{M-1}\sum_{i=1}^{M}\left(\frac{\hat{y}_i - \mu_{\hat{y}}}{\sigma_{\hat{y}}}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right),$$

where μ_y, σ_y, $\mu_{\hat{y}}$, and $\sigma_{\hat{y}}$ are the means and standard deviations for the true and estimated ages, respectively. Clearly, a larger score on this metric indicates better performance.

4. RESULTS AND DISCUSSION

In this section we summarize our results obtained with the experimental setup presented in Section 3. In the first experiment we investigated the effectiveness of pre-processing the i-vectors through LDA for age estimation, where the dimensionality is reduced from 500 to 20. The outcome of this experiment is presented in Table 1 in terms of the MAE and Pearson's correlation measure. The results are shown both individually and combined for male and female speakers in the test set. It can be seen that LDA consistently provides gains in age estimation performance. Additionally, LDA helps reduce the SVR training time from an average of 62 seconds to 6 seconds on a server machine with Intel® Xeon® CPU E5-2690. This speedup in the training can also potentially improve the scalability of the system for much larger corpora with CTS data from thousands of callers. Accordingly, we include LDA processing in all our subsequent experiments.

In the next experiment, we measured the age estimation performance with the logarithm of the age as the target in training the SVR which helps emphasize the relative regression errors as opposed to the absolute errors. Table 2 shows the age estimation results for this experiment. It is seen from the table that penalizing estimation errors for younger speakers (compared with older speakers) results in improved age estimation performance. This behavior is expected because i) as illustrated in Fig. 2, the age distribution of speakers is right-skewed, i.e., it is biased towards younger speakers, and ii) as noted previously, transforming the true age targets according to (4) has the benefit of setting a lower bound on the estimated ages.

In the next set of experiments we explored various feature representations as input to the age estimation systems with GMM-UBM based i-vectors as well as DNN senone posterior based i-vectors. Results of these experiments are summarized in Tables 3 and 4. Several observations can be made from the results presented in these tables. First, with the GMM-UBM based i-vector system, MFCC-SDC features seem to provide a slight benefit in age estimation performance, but only for male speakers. Second, age estimation performance for both MFCC-SDC and MFCC-Δ (MFCC with delta and delta-delta contextualization) is better for female speakers than male speakers. This may be attributed to the fact that female speakers are better represented in the training set (see Fig. 2). Third, an i-vector level combination of MFCC-Δ and MFCC-SDC systems results in improved performance for both female and male speakers (see last row in Table 3). Fourth, comparing the results in Table 3 and Table 4, it is clear that the age estimation system with DNN based i-vectors consistently outperforms the system with GMM-UBM based i-vectors. Fifth, using speaker-adapted fMLLR features compared with the baseline acoustic features results in further gains in the age estimation performance, in particular for male speakers. Finally, the best individual system performance is obtained with i-vectors extracted using fMLLR features and senone posteriors obtained from a DNN trained with fMLLR features concatenated with i-vectors.

Table 3. Age estimation performance with MFCC and MFCC-SDC systems as well as their i-vector level fusion. GMM-UBM based i-vector extractors are used.

	Male		Female		Both	
System	MAE	ρ	MAE	ρ	MAE	ρ
MFCC - Δ	6.5	0.79	5.7	0.87	6.0	0.84
MFCC - SDC	**6.1**	0.79	5.7	0.87	**5.8**	0.84
Combination	5.8	0.84	5.5	0.89	5.6	0.87

Table 4. Age estimation performance with MFCC, MFCC-SDC, fMLLR, and fMLLR+i-vector systems as well as their i-vector level fusion. DNN based i-vector extractors are used.

	Male		Female		Both	
System	MAE	ρ	MAE	ρ	MAE	ρ
MFCC - Δ	5.3	0.87	5.1	0.90	5.2	0.89
MFCC - SDC	5.4	0.85	5.1	0.90	5.2	0.88
fMLLR	4.9	0.89	5.0	0.90	5.0	0.89
fMLLR/i-vector	**4.7**	0.89	**4.7**	**0.91**	**4.7**	**0.91**
Combination	4.8	0.90	4.7	0.92	4.8	0.91

5. CONCLUSION

In this paper we presented an automatic age estimation system based on DNN senone posterior i-vectors and SVR modeling. It was shown that the use of the phonetically-aware i-vector extractor, compared with the GMM-UBM based counterpart, could improve the age estimation performance. It was demonstrated that processing i-vectors through an LDA transform trained with discrete age labels not only improved the performance, but also dramatically sped-up the SVR training process. Further improvements in the age estimation performance were achieved by employing the use of the logarithm of the age as the target in training the SVR to further penalize estimation errors for younger speakers compared with older speakers. The system proposed in this study achieved solid performance on the NIST SRE 2010 telephony test set.

6. ACKNOWLEDGEMENTS

The authors wish to thank Mohamed Omar of IBM Research for helpful discussions.

7. REFERENCES

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language–state-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 4 – 39, 2013.

[2] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. A. Müller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel, "Comparison of four approaches to age and gender recognition for telephone applications," in *Proc. IEEE ICASSP*, Honolulu, HI, April 2007, pp. 1089–1092.

[3] T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth, "Age and gender recognition based on multiple systems - early vs. late fusion," in *Proc. INTERSPEECH*, Makuhari, Japan, September 2010, pp. 2830–2833.

[4] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 151 – 167, 2013.

[5] G. Dobry, R. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 1975–1985, September 2011.

[6] M. H. Bahari, M. McLaren, H. V. hamme, and D. A. van Leeuwen, "Age estimation from telephone speech using i-vectors," in *Proc. INTERSPEECH*, Portland, OR, September 2012, pp. 506–509.

[7] ——, "Speaker age estimation using i-vectors," *Eng. Appl. of AI*, vol. 34, pp. 99–108, 2014.

[8] A. Fedorova, O. Glembek, T. Kinnunen, and P. Matějka, "Exploring ANN back-ends for i-vector based speaker age estimation," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 3036–3040.

[9] S. Schötz and C. Müller, "A study of acoustic correlates of speaker age," in *Speaker Classification II*, ser. Lecture Notes in Computer Science, C. Müller, Ed. Springer Berlin Heidelberg, 2007, vol. 4441, pp. 1–9.

[10] W. Spiegl, G. Stemmer, E. Lasarcyk, V. Kolhatkar, A. Cassidy, B. Potard, S. Shum, Y. C. Song, P. Xu, P. Beyerlein, J. D. Harnsberger, and E. Nöth, "Analyzing features for automatic age estimation on cross-sectional data," in *Proc. INTERSPEECH*, Brighton, UK, September 2009, pp. 2923–2926.

[11] M. Kockmann, L. Burget, and J. Cernocký, "Brno university of technology system for Interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, Makuhari, Japan, September 2010, pp. 2822–2825.

[12] S. Safavi, M. J. Russell, and P. Jancovic, "Identification of age-group from children speech by computers and humans," in *Proc. INTERSPEECH*, Singapore, Singapore, September 2014, pp. 243–247.

[13] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, August 1980.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[15] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 1695–1699.

[16] M. K. Omar and J. Pelecanos, "Training universal background models for speaker recognition," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2010)*, Brno, Czech, June 2010, pp. 52–57.

[17] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic press, 1990.

[18] V. Vapnik, *The nature of statistical learning theory*. New York: Springer, 1995.

[19] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[20] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora," in *Proc.INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 950–953.

[21] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Mixer 6," in *Proc. LREC*, Valletta, Malta, May 2010.

[22] B. Bielefeld, "Language identification using shifted delta cepstrum," in *Proc. 14th Annual Speech Research Symposium*, Baltimore, MD, 1994.

[23] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. INTERSPEECH*, Denver, CO, September 2002, pp. 89–92.

[24] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 197–200, 2013.

[25] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. Workshop Human Lang. Tech.*, March 1994, pp. 307–312.

[26] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proc. LREC*, Lisbon, Portugal, May 2004.

[27] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 357–366, September 1995.

[28] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[29] S. Ganapathy, S. Thomas, D. Dimitriadis, and S. Rennie, "Investigating factor analysis features for deep neural networks in noisy speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 1898–1902.